HETEROGENEOUS FEATURE SELECTION FOR CLASSIFICATION OF CUSTOMER LOYALTY FAST MOVING CONSUMER GOODS (CASE STUDY: INSTANT NOODLE)

[1]HENI SULISTIANI, [2]ARIS TJAHYANTO

[1]Department of Information System, STMIK Teknokrat Lampung, Indonesia

[2] Department of Information System, Institut Teknologi Sepuluh Nopember, Indonesia

E-mail: [1]henie.tekno@gmail.com, [2]atjahyanto@gmail.com

ABSTRACT

In the face of ASEAN open market, the actors (Fast Moving Consumer Goods industry) must increasingly explore patterns of business development because of tough competition and challenges in the market. One of strategy for surviving in high competition is retain the customer loyalty. The data were usually obtained from various sources and contains heterogeneous features, such as numerical and non-numerical features. The datasets with heterogeneous features can affect feature selection results that are not appropriate because of the difficulty of evaluating heterogeneous features concurrently. In this paper, we propose a method that combine the features transformation and subset selection based on mutual information to obtain feature subset that able to improve performance classification algorithm. Analysis comparative among before feature subset selection, dynamic mutual information (DMI) methods, p-value methods and researcher estimate were also done. Feature transformation (FT) is another way to handle the selection of heterogeneous features. The datasets were obtained from the survey of customers fast moving consumer goods, with a total of 386 respondents. By applying unsupervised feature transformation and dynamic mutual information methods, can be known the relevant features that affected the performance of decision tree algorithm. The accuracy and F-measure increased of the DMI- unsupervised-feature-transformation compared to all features (without features subset selection), p-value methods and manual features subset selection. The accuracy and F-measure for DMI-unsupervised-feature transformation are 76.68% and 73.5% respectively.

Keywords: *Classification, Customer Loyalty, Decision Tree, Feature Transformation, Mutual Information*

1. INTRODUCTION

The Indonesia Kantar Worldpanel report shows an increase of 14% of 2012 to 2013 sales of products Fast Moving Consumer Goods (FMCG) in the whole Indonesia territory, both in urban and rural [1]. In the face of ASEAN open market, the actors (FMCG industry) must increasingly explore patterns of business development because of strict competition and challenges in market [2]. Therefore required to determine the right marketing strategy in order to retain the customers and able to survive in this market environment. One of appropriate marketing strategies for surviving in high competition is retain customer loyalty [3].

The machine learning approaches and data mining techniques can be used for predicting customer loyalty. The success of information discovery in data processing is influenced by several factors. One key factor is the quality of data [4]. If the data has too much noise, or a lot of data that is redundant and irrelevant, the training process of information discovery would have difficulty. In the pre-process phase, the features selection is one of the important parts [5] in finding an optimal feature subset, removes irrelevant or redundant feature [6]. Further, well-chosen features can improve classification accuracy substantially, or equivalently, reduce the amount of training data needed to obtain a desired level of performance [7]. Often, when the mining process are presented with a number of attributes that are not small, many not useful attributes that are used for prediction.

Data are usually obtained from various sources and contains features heterogeneous, i.e. numerical and non-numerical features. Datasets with heterogeneous features can affect feature selection results that are not appropriate because of the difficulty of evaluating heterogeneous features concurrently [8]. Some of the methods proposed for selection heterogeneous features. One of them, research on regression analysis was used to identify important variables that influence buying behavior of customers in service companies and using Markov chains to model the probability of transition from behavioral changes [9]. However, regression analysis has the disadvantage i.e. difficult to interpret the intercept coefficient and can lead to the interpretation that is inconsistent with the actual conditions [10].

Discretization method for the selection heterogeneous features was proposed by [11], this method of dividing the values of numerical features into multiple intervals and represent them with a various of non-numeric values. However, discretization method often leads to loss of information because it reduces the distance and the order in original numerical features [12, 13]. Feature transformation is another way to handle the feature subset selection with unify the format of data sets and can be used in a traditional feature selection algorithm.

Feature transformation method for handling heterogeneous data using unsupervised feature transformation (UFT) was proposed by [8, 14], introduced the least information distortion and is more reliable because the processed data are bias-free. In this paper, we propose a method to combine the features transformation and feature subset selection based on mutual information. Dynamic feature selection method of mutual information [15] is done by calculating the relationship between the label feature class. This method is able to minimize the distortion of information, measure various types of relationships including stronger against the nonlinear and MI robust to feature that contains noise.

This paper is organized as follows. Section 2 describes research method of the proposed methods. Section 3 describes the results and analysis. Section 4 concludes this study.

2. RESEARCH METHOD

The datasets used in this study is result the survey of customers fast moving consumer goods, with a total of 386 respondents. This datasets contains 5 numerical feature and 21 non-numerical features (as shown in Table 1). This dataset has two classes of data, they are 'yes' for the customers loyal dan 'no' for customers are not loyal. The stages of methods in this study are transformation from heterogeneous feature to homegeneous feature using unsupervised feature transformation (UFT), feature subset selection using dynamic mutual information (DMI), classification using decision tree algorithm, performance measurement and analysis comparative among before selection feature, DMI methods, p-value methods and researcher estimate. General description of the research methods is shown in Fig. 1.

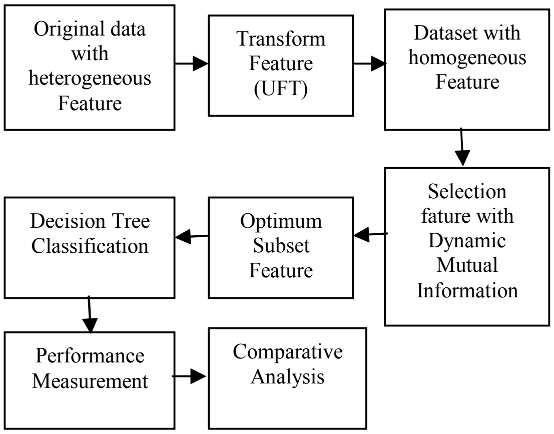
Original data with heterogeneous Feature
Transform Feature (UFT)
Dataset with homogeneous Feature
Selection fature with Dynamic Mutual Information
Optimum Subset Feature
Decision Tree Classification
Performance Measurement
Comparative Analysis

Figure 1. The Research Methodology

Table 1 Description of Datasets

Feature	Description	Type
f1	Gender	Non-numeric
f2	Age	Numeric
f3	Address	Non-numeric
f4	Marital status	Non-numeric
f5	Status of residence	Non-numeric
f6	Job status	Non-numeric
f7	Education	Non-numeric
f8	Brands	Non-numeric
f9	Point of purchase	Non-numeric
f10	Promotion media	Non-numeric
f11	Reason to consumption	Non-numeric
f12	Duration of consumption	Non-numeric
f13	Distance purchases	Non-numeric
f14	Number of buying	Non-Numeric
f15	Consumption average	Numeric
f16	Display products	Non-Numeric
f17	Satisfaction of price	Non-Numeric
f18	Brands satisfaction	Non-Numeric
f19	Recomendation	Non-Numeric
f20	Comment	Non-Numeric
f21	Expenses	Numeric
f22	Number of consumption per once	Numeric
f23	Switch brands	Non-Numeric
f24	Reason to switch	Non-Numeric
f25	The other brands consumption	Numeric
f26	Behavior back	Non-Numeric

2.1. Unsupervised Feature Transformation (UFT)

This process is intended to change the non-numeric features into numeric features that depends on the original non-numeric features itself. UFT is derived from an analytical relationship between mutual information (MI) and entropy. The purpose of UFT is to find a numerical *X'* to substitute the original non-numerical feature *X*, and *X'* is constrained by *I*(*X'*;*X*) = *H*(*X*). This constraint makes the MI between the transformed *X'* and the original *X* to be the same as the entropy of the original *X*. This condition is critical because it ensures that the original feature information is preserved, when non-numerical features are transformed into numerical features. It is also worth nothing that the transformation is independent of class label, so that the bias introduced by class label can be reduced. After it is processed by UFT methods, the datasets's format which have heterogeneous features can be combined to numerical features entirely. The solution for UFT methods is [8, 14]:

$$\mu_i^* = \left[(n\text{-}i)\text{-}\sum_{k=1}^{i}(n\text{-}k)p_k\right]\sqrt{(1\text{-}\sum_i p_i^3)/\sum_{i\neq j} p_i\, p_j (i\text{-}j)^2} \qquad (1)$$

where

$$\boldsymbol{\sigma_i^* = p_i \qquad i \,\epsilon\, \{1, \dots, n\}}$$

Based on the solution, equation (1), UFT methods can be formalized as (figure 2).

```
Algorithm: UFT
Input: dataset D, which heterogeneous feature f_j, j ϵ {1, ..., m}
Output: transformed dataset D' with pure numerical features
for j = 1 to m do
    if feature f_j is non-numerical then
        n = size(unique(f_j));
        {s_i|i = 1, ..., n} is the set of non-numerical values in feature f_j
        p_i is the probability of s_i
        for i = 1 to n do
            $\mu_i = [(n-i) - \sum_{k=1}^{i}(n-k)p_k]\sqrt{(1-\sum_i p_i^3)/\sum_{i\neq j} p_i\, p_j (i-j)^2}$;
            $\sigma_i = p_i$ ;
            use Gaussian distribution $\mathcal{N}(\mu_i, \sigma_i)$ to generate numerical data and substitute the values equal to $s_i$ in feature $f_j$
        end for
    end if
end for
```

Figure 2. UFT Algorithm

2.2. Information Theory

In information theory, entropy is a key measure of information. Since it is capable of quantifying the uncertainty of random variables and scaling the amount of information shared by them effectively, it has been widely used in many fields. In the case of feature selection, relevant features have important information relating to the class, otherwise irrelevant features contain little information relating to class [16, 17]. To measure the uncertainty of a random variables X with discrete values can use entropy function *H*(*X*), which is defined as:

$$H(X) = -\sum_{x\in X} p(x) l\text{c} \quad (x) \qquad (2)$$

Where *p*(*x*)=Pr(*X*=*x*) is the probability density function of *X*. The *joint entropy H*(*X*, *Y*) of *X* and *Y* is

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) \qquad (3)$$

To measure show much information is shared by two variables *X* and *Y*, a concept termed *mutual information I*(*X*; *Y*) is defined as:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y) l\text{c} \quad \frac{p(x,y)}{p(x)p(y)} \qquad (4)$$

If value of *I*(*X*; *Y*) is very high, then *X* and *Y* are closely related with each other; otherwise, if *I*(*X*; *Y*) ≤ 0, then two variables are totally unrelated.

2.3. Dynamic Mutual Information

In this section a feature subset selection algorithm based on dynamic mutual information. Generally, instances in *T=D(F, C)* can be classified into two kinds: labeled and unlabeled. Suppose *S* is the subset of already selected features and the instances *D* has been partitioned into two parts with respect to the labels *C*, i.e., labeled instances $D_l \subseteq D$ and unlabeled instances $D_u \subseteq D$, where $D_l \cap D_u = \emptyset$ and $D_l \cup D_u = D$. And then pick the candidate feature *f* out from *F*, which will classify as more instances in D_u as possible. Feature with the largest mutual information on D will be selected.

Before identifying feature on each stage, first step of algorithm is estimates the value of mutual information from candidate features *f* with label class *C*. This recursion procedure is similar to the first step of the other mutual information algorithm, which takes the value of mutual information feature with the highest estimated on the whole sample space *D*. During the calculation phase, features (*fi*) which has a mutual information value as big as zero (0) then it will be removed from *F* and features that have the highest value of mutual information will be selected. The purpose of this step is to prevent them from being recalculated in estimating mutual information in next time. After doing that, the algorithm will go to the next round to pick up other candidate features. This procedure will be repeated, until there has no candidate features in *F* or the number of the unlabeled instances is equal to inconsistency count of *T*.

The search strategy used is sequential forward search. For each iteration, the value of the mutual information *I(fi, C)* will be calculated for all the features in candidate *F*. Explicitly, the details of algorithms as shown in Fig. 3.

```
Algorithm 1. DMIFS: feature selection using dynamic mutual information.
Input: A training dataset $T=D(F,C)$
Output: Selected features $S$.
Initialize relative parameters: $F = F$;
    $S = \emptyset$; $D_u = D$; $D_l = \emptyset$;
Repeat
  For each feature $f \in F$ do
    Calculate its mutual information $I(C;f)$ on $D_u$;
    If $I(C;f) = 0$ then $F = F - \{f\}$;
  Choose the feature $f$ with the highest $I(C;f)$;
  $S = S \cup \{f\}$; $F = F \backslash \{f\}$;
  Obtain new labeled instances $D_l$ from $D_u$ induced by $f$;
  Remove them from $D_u$, i.e., $D_u = D_u \backslash D_l$;
Until $F = \emptyset$ or $|D_u| = I_T$
```

Figure 3. Dynamic Mutual Information Algorithm

2.4. C4.5 Decision Tree Classifier

The C4.5 algorithm was introduced by [18] as an improved version of ID3. The improvements that distinguishes C4.5 algorithm from ID3 are the ability to handle the numeric type feature, perform pruning of decision tree and deriving a set of rules. The C4.5 decision tree learning is one of the most widely used and practical methods for inductive inference [19]. The C4.5 algorithms uses the criteria gain in determining the features that become nodes on the tree-induced. The construction begins at the root node where each attribute is evaluated using a statistical test to determine how well it can classify the training samples [20].

The best attribute is chosen as the test at the root node of the tree. A statistical test used in C4.5 for assigning an attribute to each node in the tree also employs an entropy-based measure. The assigned attribute is the one with the highest information gain ratio among attributes available at that tree construction point. The information gain ratio GainRatio(A, S) of an attribute A relative to the sample set S is defined in Eq.5 as follow:

$$G \qquad = \frac{G \quad (A,S)}{S\imath \qquad (A,S)} \tag{5}$$

where

$$G \quad (A,S) = E \quad (S) - \sum_{a \in A} \frac{|S_a|}{|S|} E \quad (S_a) \tag{6}$$

and

$$S\imath \qquad (A,S) = -\sum_a \frac{|S_a|}{|S|} lc \;\; {}_2 \frac{|S_a|}{|S|} \tag{7}$$

Where S_a is the subset of S for which the attribute A has the value a.

3. RESULT AND ANALYSIS

The classification prediction model was processed using Weka and validated by 10-fold cross validation method. This paper conducted two test scenarios. The first scenario discusses the features transformation into a homogeneous feature and feature subset selection most associated with the class label in the analysis of customer loyalty fast moving consumer goods based on the mutual information between feature and class label. The second scenario is classifying the customer loyalty using a decision tree:

a. Use all the features of the original data (before feature subset selection);
b. Use the selected features by the dynamic mutual information (DMI) method;
c. Use the selected features by the p-value method; and
d. Using the selected features on the basis of researcher estimation.

The combination of these features were used in building a decision tree classification of customer loyalty. Based on test scenarios described above, then perform a comparative analysis of the accuracy level for the classification of customer loyalty. The performance evaluations were performed using the value of accuracy and F-measure. The results of rank from the first test scenarios presented in Table 2.

Table 2. The rank of Features Subset Selection

Feature Subset Selection	Selected Features
Before features selection	f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15, f16, f17, f18, f19, f20, f21, f22, f23, f24, f25, f26
Dynamic mutual information	f21, f15, f2, f3, f24, f25, f11, **f7**, **f6**, **f8**, f14, f13, f12, f10, f17, f22, f18, f5, f9, f23, f4, f26, f19, f20, f1, f16
Manual features selection	f2, f25, f3, f4, **f6**, f15, f21, **f7**, **f8**, f10, f11, f18, f12, f13, f24, f1, f23, f26, f5, f16, f9, f17, f14, f22, f19, f20
P-value	**f6**, f26, f17, f10, f1, **f8**, **f7**, f24, f16, f12, f11, f14, f3, f18, f5, f13, f25, f21, f19, f9, f15, f20, f4, f22, f2, f23

The rank result was obtained using the method of dynamic mutual information, p-value and researcher estimation. There are three features that are always selected, if the number of the features is limited to ten features. These features are job status (f6), education (f7) and brands (f8). It is means that the three of these features affected to customer loyalty.

The testing of performance classification comparison was conducted to determine the performance of the classifier, which generated predictions with the smallest error value. Figure 4 shows the comparison of accuracy from several method of feature subset selection using predictive models decision tree algorithm.

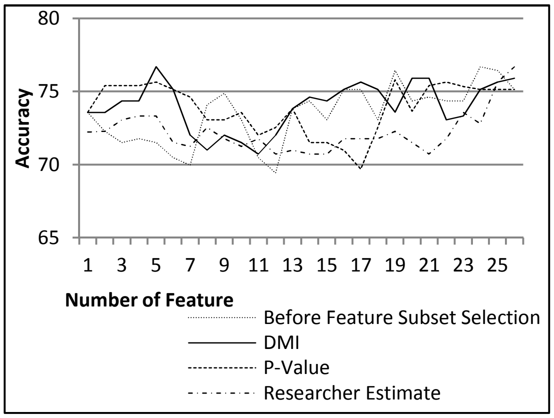

Figure 4. Comparison Accuracy of Decision Trees Algorithm

Based on Figure 4 can be infered that using the dynamic mutual information method gave better accuracy when compared to a p-value method, researcher estimation and using all the features before feature subset selection in the use of the number of features as many as five. Whereas on Fig. 5 shows the a comparison of F-measure for feature selection method based on predictive models of decision trees algorithm.

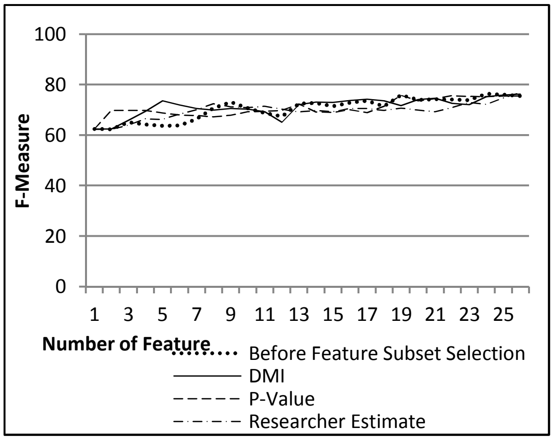

Figure 5 Comparison of F-Measure Decision Tree Algorithm

Figure 5 depicts that the method of dynamic mutual information has difference value of F-measure that significant in the use of the number of features as many as five. The highest classification performance of decision tree method that applied feature subset selection dynamic mutual information on the five features selected, i.e. the value of accuracy 76.68% and the value of F-measure 73.5%. By applying the DMI-unsupervised-feature transformation methods, it can be seen that the relevant features affected the performance of decision tree algorithm, that are consumption average (f15), age (f2), expenses (f21), reason to switch (f24) and address (f3). From a number of 386 data at five selected features, the 296 data (76.68%) can be classified correctly, whereas 90 data (23.32%) were incorrectly classified (as shown in Table 3).

Table 3. Confusion Matrix Decision Tree Algorithm

Test Result	Customer Loyalty	
	Loyal	Not Loyal
Positive	266	18
Negative	72	30

Table 3 is confusion matrix for amount of data test of estimated loyal customers is 284. A number of 266 customers is predicted correctly as loyal customer (true-positive/ TP), whereas 18 customers predicted incorrectly (false-positive / FP) by the classifier of decision tree. The testing on the non-loyal customers yield 30 customers (true-negative/ TN) predicted correctly and customers are not loyal, otherwise 72 customers (false-negative/ FN) predicted incorrectly as a not loyal customer.

4. CONCLUSION

The unsupervised feature transformation was used to change the features of non-numeric into a numeric feature based on the rule of Gaussian distribution. Whereas the dynamic mutual information method was used for features subset selection based on value of mutual information between features and class label. By combining both methods, we show that we can improve the performance of decision tree algorithms in classification of customer loyalty on fast moving consumer goods. The best performance of this research was obtained using five features in the top rankings.

By applying DMI-unsupervised-feature transformation methods, it can be seen that the relevant features affected the performance decision tree algorithm, namely are consumption average (f15), age (f2), expenses (f21), reason to switch (f24) and address (f3). The accuracy and F-measure for DMI-unsupervised-feature transformation values respectively are 76.68% and 73.5% respectively

REFRENCES:

[1] Sundari, C, "Mengenal Fast Moving Consumer Goods", June 2014, *Kompasiana: http://www.kompasiana.com* (accessed Ocktober 15, 2015).

[2] Kurniawan, "Kantar Worldpanel Pertumbuhan Industri FMCG Indonesia Tertinggi di Asia", October 16, 2014, *Gatra: http://www.gatra.com* (accessed October 31, 2015).

[3] Santoso, Teguh Budi, "Analisa dan Penerapan Metode C4.5 untuk Prediksi Loyalitas Pelanggan", *Jurnal Ilmiah Fakultas Teknik LIMIT'S Vol. 10 No.1*, 2012.

[4] Purbasari, I. Y., & B. Nugroho, "Bencmarking Algoritma Pemilihan Atribut Pada Klasifikasi Data Mining", *SNASTIA*, 2013, pp. 47-54.

[5] Blum, Avrim L., & Pat Langley, "Selection of relevant features and examples in machine learning", *Elsevier Science Artificial Intelligence*, 1997, pp. 245-271.

[6] Tjahyanto, Aris, Yoyon K Suprapto, and Diah P Wulandari, "Spectral-based Features Ranking for Gamelan Instruments Identification using Filter Techniques", *TELKOMNIKA*, Vol. 11, No. 1, March 2013, pp. 95-106.

[7] Forman, George, "An Extensive Empirical Study of Feature Selection for Text Classification", *Journal of Machine Learning Research 3*, 2003, pp. 1289-1305.

[8] Pawening, Ratri Enggar, Tio Darmawan, Rizqa Raaiqa Bintana, Agus Zainal Arifin and Darlis Herumurti, "Feature Selection Methods Based on Mutual Information for Classifying Heterogeneous Features", *Journal of Computer Science and Information, Volume 9, Issue 2*, June 2016, pp. 106-112.

[9] Cheng, C.-J., Chiu, S., Cheng, C.-B., & Wu, J.-Y, "Customer lifetime value prediction by a Markov chain based datamining model: Application to an auto repair and maintenance company in Taiwan", *Transactions E: Industrial Engineering*, 2012, pp. 849–855.

[10]Novita, M, "Regresi Linier Sederhana", *Academia:http://www.academia.edu/4378028/regresi_linier_sederhana* (accessed November 2, 2015)

[11]Liu, H., & R Setiono, "Feature Selection via Discretization", *IEEE Transactions on Knowledge and Data Engineering, vol. 9, no. 4*, 1997, pp. 642-645.

[12]Hu, Q, D. Yu, J. Liu, J. and C. Wu, "Neighborhood Rough Set Based Heterogeneous Feature Subset Selection", *Information Science 178*, 2008, pp. 3577–3594.

[13]Chong, J. Y., & A. K. Wong, "Class-Dependent Discretization for Inductive Learning from Continuous and Mixed-Mode Data", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995, pp. 641-651.

[14]Wei, M., T. W. Chow, and R. H. Chan, "Heterogeneous Feature Subset Selection usng Mutual Information-Based Feature Transformation", *Elsevier Neurocomputing*, 2015, pp. 1-13.

[15]Liu, Huawen, Jugui Sun, Lei Liu, and Huijie Zhang, "Feature Selection with Dynamic Mutual Information", *Elsevier Pattern Recognition 42*, 2009, pp. 1330-1339.

[16] Li, Y., M. Xie, & T. Goh, "A Study of Mutual Information Based Feature Selection for Case Based Reasoning in Software Cost Estimation", *Expert System with Application*, 2009, pp. 5921-5931.

[17]Kwak, N., & C.-H Choi, "Input Feature Selection by Mutual Information Based on Parzen Window", *IEEE Transactional Pattern*

Analitycal Matchematics Intelligence, 2002, pp. 1667-1671.

[18]Quinlan, J. R., “C4.5: Programs for Machine Learning”, *Machine Learning*, 1994, pp. 235-240.

[19]Polat, Kemal and Salih Gunes, “A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems”, *Science Direct Expert Systems with Applications* 36, 2009, pp. 1587–1592.

[20]Setsirichoka, Damrongrit, Theera Piroonratanaa, Waranyu Wongsereea, Touchpong Usavanaronga, Nuttawut Paulkhaolarnb, Chompunut Kanjanakornb, Monchan Sirikongb, Chanin Limwongsec and Nachol Chaiyaratana, “Classification of Complete Blood Count and Haemoglobin Typing Data by a C4.5 Decision Tree, a Naive Bayes Classifier and a Multilayer Perceptron for Thalassaemia Screening”, *Science Direct Biomedical Signal Processing and Control 7*, 2012, pp. 202-212.